Extending the entry consistency model to enable efficient visualization for code-coupling grid applications

Gabriel Antoniu, Loïc Cudennec, Sébastien Monnet

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Extending the entry consistency model to enable efficient visualization for code-coupling grid applications

Gabriel Antoniu, Loïc Cudennec, Sébastien Monnet

N°5813

Janvier 2006

Systèmes numériques

ISSN 0249-6399 ISRN INRIA/RR--5813--FR+ENG

Extending the entry consistency model to enable efficient visualization for code-coupling grid applications

Gabriel Antoniu*, Loïc Cudennec†, Sébastien Monnet‡

Systèmes numériques
Projet Paris

Rapport de recherche n°5813 — Janvier 2006 — 17 pages

Abstract: This paper addresses the problem of efficient visualization of shared data within code coupling grid applications. These applications are structured as a set of distributed, autonomous, weakly-coupled codes. We focus on the case where the codes are able to interact using the abstraction of a shared data space. We propose an efficient visualization scheme by adapting the mechanisms used to maintain the data consistency. We introduce a new operation called *relaxed read*, as an extension to the *entry consistency* model. This operation can efficiently take place without locking, in parallel with *write* operations. On the other hand, the user has to relax the consistency constraints, and accept slightly older versions of the data, whose "freshness" can however still be controlled. This approach has been implemented within the JuxMem grid data-sharing service, and its efficiency is clearly demonstrated by our preliminary experimental results.

Key-words: JuxMem, Consistency, Visualization, Grid, Data Sharing.

(Résumé : tsvp)

This paper has been submitted to the 6th IEEE International Symposium on Cluster Computing and the Grid (CCGrid06).

* Gabriel.Antoniu@irisa.fr
† Loic.Cudennec@irisa.fr
‡ Sebastien.Monnet@irisa.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Extension du modèle de cohérence à l'entrée pour la visualisation dans les applications de couplage de codes sur grilles

Résumé : Ce papier s'intéresse au problème de la visualisation des données partagées dans les applications à base de couplage de codes sur les grilles. Nous proposons d'améliorer l'efficacité de la visualisation en intervenant sur les mécanismes de gestion des données répliquées et plus particulièrement au niveau du protocole de cohérence. La notion de *lecture relâchée* est alors introduite comme une extension du modèle de cohérence à l'entrée (*entry consistency*). Ce nouveau type d'opération peut être réalisé sans prise de verrou, en parallèle avec des écritures. En revanche, l'utilisateur *relâche* les contraintes sur la *fraîcheur* de la donnée et accepte de lire des versions *légèrement* anciennes, dont le retard est néanmoins contrôlé. L'implémentation de cette approche au sein du service de partage de données pour grilles JuxMem montre des gains considérables par rapport à une implémentation classique basée sur des lectures avec prise de verrou.

Mots-clé : JuxMem, Cohérence, Visualisation, Grille, Partage de données.

1 Introduction

With the growing demand of computing power, grid computing [12] has emerged as an appealing approach, allowing to federate and share computing and storage resources among multiple, geographically distributed sites (universities, companies, etc.). Thanks to this aggregated computing power, grids are typically useful to solve computationally intensive, parallel and/or distributed applications. In most cases, grids consist of a hierarchical federation of clusters. Often, SANs's, such as Giga Ethernet or Myrinet are used to connect nodes within a given cluster. The various clusters may be interconnected through a higher-latency network, which can be a dedicated WAN whose bandwidth may reach 1 Gb/s or more.

A particular class of applications running on grids relies on the *code-coupling* paradigm: such an application is designed as a set of (usually) parallel codes, each of which runs on a different cluster. The computation is distributed in such a way that transfers between clusters are minimized. However, some data and synchronization messages still have to be exchanged among the clusters.

Code-coupling is used in high-performance computing. These computations can be very long, and it is generally impractical to wait for the end of the application to see if the results are correct. In order to see the progress of the application, it is often useful to have the ability to perform an efficient visualization of the running process, without degrading the overall performance of the computation. To allow the state of the computation to be monitored, pieces of data shared by different codes need to be accessed.

In grid environments, as in other distributed systems, data sharing is a crucial issue. Currently, the most widely-used approach relies on the *explicit data access model*, where clients have to move data to computing servers. A typical example is the use of the GridFTP protocol [3]. Though this protocol provides authentication, parallel transfers, checkpoint/restart mechanisms, etc., it is still a transfer protocol which requires *explicit* data localization by the programmer. Such a low-level approach makes data management on grids rather complex. On the other hand, the concept of *transparent data access* in distributed systems through the illusion of a shared memory has intensively been studied in the context of distributed shared memory systems (DSM) since the late eighties ([13, 11, 4, 10]). Nevertheless, DSM systems have been designed to address small scale physical architectures, usually made of tens (up to a hundred) of nodes and have usually been used on clusters. Furthermore, most of the data consistency models and protocols assume that the infrastructure is *static, without failures*. For instance, they often implicitly assume stable entities. These hypotheses are not longer valid within the grid context, where failures are part of the systems' properties. Therefore, *fault tolerance* and *volatility* increase the difficulty of designing a system providing transparent data access. The predominance of grid systems based on *explicit* transfers (GridFTP [3],

IBP [9], etc.) demonstrates that transparent data sharing upon large scale architectures stays a real challenge.

In order to overcome these limitations and make a step forward towards a real virtualization of the management of large-scale distributed data, the concept of *grid data-sharing service* has been proposed [5]. The idea is to provide *transparent access* to distributed grid data: in this approach, the user accesses data via global identifiers. The service which implements this model handles data localization and transfer without any help from the programmer. It transparently manages data persistence in a dynamic, large-scale, distributed environment. The data sharing service concept is based on a hybrid approach inspired by Distributed Shared Memory (DSM) systems (for transparent access to data and consistency management) and peer-to-peer (P2P) systems (for their scalability and volatility-tolerance). The JuxMem (Juxtaposed Memory) platform [5] (described in more detail in Section 2) illustrates the grid data-sharing concept. JuxMem relies on JXTA [1], a generic P2P software platform initiated by Sun Microsystems. JuxMem also serves as an experimental framework for fault-tolerance strategies and data consistency protocols.

In this paper, we focus on the problem of efficient data visualization within code-coupling applications designed for grid architectures. The goal is to modify the data consistency protocol behavior in order to efficiently support the presence of a visualization process (that we call *observer*). This paper proposes an *extension* of the *entry consistency* model and a corresponding protocol that allows efficient reads, *possibly concurrent* with writes to a given data. As a counterpart, the observer has to relax the consistency constraints, and accept slightly older versions of the data, whose “freshness” can however still be controlled.

The next Section introduces the JuxMem grid data sharing service. Section 3 briefly describes the consistency model and explains the proposed protocol extensions. An experimental evaluation is presented in Section 4. Finally, Section 5 discusses the contribution and the future work.

2 JuxMem : A decoupled architecture combining data consistency and fault-tolerance

2.1 JuxMem overview

To experiment our approach, we have used the JuxMem software experimental platform for grid data sharing, described in [6, 5]. From the user’s perspective, JuxMem is a service providing transparent access to persistent, mutable shared data.

JuxMem has a *hierarchical* software architecture, which mirrors a hardware architecture consisting of a federation of distributed clusters. Figure 1 shows the hierarchy of the entities

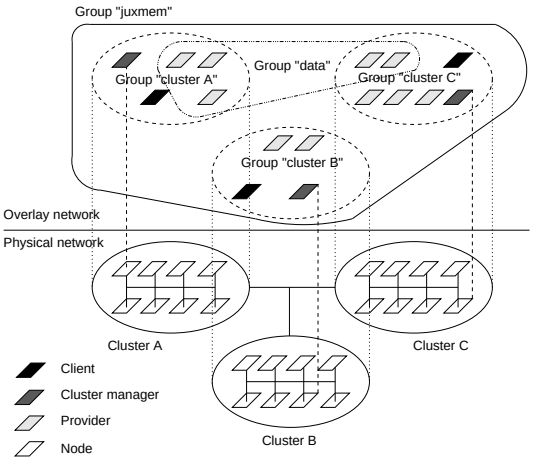


FIG. 1 – Hierarchy of the entities in the network overlay defined by JuxMem.

defined in JuxMem, consisting of a network of peer groups (`cluster` groups A, B and C on the figure), which usually correspond to clusters at the physical level. All the groups belong to a wider group, which includes all the peers which run the service (the `juxmem` group).

Each `cluster` group includes several kinds of nodes. Those which provide memory for data storage are called *providers*. Within each `cluster` group, the available providers are managed by a node called *cluster manager*. Finally, a node which simply uses the service to allocate and/or access data blocks is called *client*. It should be stressed that a node may at the same time act as a cluster manager, a client, and a provider. However, for the sake of clarity, each node only plays a single role on the figure.

When allocating memory, the client has to specify on how many clusters the data should be replicated, and on how many nodes in each cluster. This results into the instantiation of a set of data replicas, associated to a group of peers called `data` group. The allocation primitive returns a global data ID, which can be used by the other nodes to identify existing data. To obtain read and/or write access to a data block, the clients only need to use this ID.

The `data` group is also hierarchically organized, as illustrated on Figure 2: the *Global Data Group (GDG)* gathers all provider nodes holding a replica of the same piece of data. These nodes can be distributed in different clusters, thereby increasing the data availability if faults occur. The GDG group is divided into *local data groups (LDG)*, which correspond to data copies located in a same cluster.

In order to access a piece of data, a client has to be attached to a specific LDG. Then, when the client performs the read/write and synchronization operations, the consistency

protocol layer manages data synchronization and data transmission between clients, LDGs and GDG, within the strict respect of the consistency model.

2.2 Starting point: a hierarchical, fault-tolerant consistency protocol

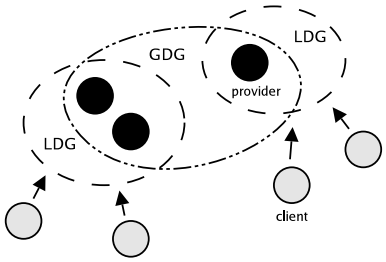


FIG. 2 – JuxMem : a hierarchical architecture.

To guarantee data consistency, JuxMem provides a *hierarchical, fault-tolerant* consistency protocol that implements the entry consistency model. The entry consistency model was first introduced in the Midway system [10]. As opposed to other relaxed models, it requires an explicit association of data to synchronization objects. This allows the model to leverage the relationship between a synchronization object that protects a critical section, and the data accessed within that section. A node's view of some data becomes up-to-date only when the node enters the associated critical section. This eliminates unnecessary traffic, since only nodes that declare their intention to access data will get updated, and only the data which will be accessed will be updated. Such a concern for efficiency makes this model a good candidate in the context of scientific grid computing.

When using the entry consistency model, exclusive accesses to shared data have to be explicitly distinguished from non-exclusive accesses by using two different primitives: `acquire`, which grants mutual exclusion; `acquireRead`, which allows non-exclusive accesses on multiple nodes to be performed in parallel.

JuxMem implements a hierarchical, home-based protocol for entry consistency, where the role of the home is played by the LDG at cluster level and by the GDG at global level. This protocol is described in detail in [8]. When using this protocol, if a client asks for a data access, its request may go through each level of the `data` group hierarchy, in order to be satisfied. For instance, when a client needs to acquire the read-lock, it sends a request to its associated LDG. If the LDG does not already have the read-lock, the LDG sends a request to the GDG. Then the lock is sent back from the GDG to the LDG and finally to the client. In this model, if a client owns a lock, its associated LDG owns the same lock.

Finally, the consistency protocol gives the priority to writers: a writer only has to wait that previous requests are satisfied, whereas a reader has to wait that no writer is asking for the lock. This strategy can cause readers starvation if two or more writers get alternatively the lock, postponing data access to readers.

3 Efficient visualization through concurrent reads and writes

3.1 Proposed enhancement: relaxed reads

We consider a scenario where an observer node reads a shared data for visualization purpose. The reads performed by this node should be efficient and low intrusive. The first idea is to take advantage of the data copies located on the client node or on its associated LDG. This provides the ability to use a data copy that is already on the client node or to fetch one from a close node (within the same cluster). The second idea is to perform the read operation without acquiring a lock. This particular read operation provides the ability to have concurrent reads and writes as it does not lock the data.

The entry consistency model guarantees that the data is up-to-date only if the associated lock has been acquired. If the associated lock has not been acquired, no guarantees are provided. The approach highlighted in this paper proposes to enable relaxed reads (i.e. without acquiring a lock) for which the user application is able to keep control on the data "*freshness*". This implies that the consistency protocol implementing this extended model respects bounds on the difference between the version of the data returned by the *rlxread* primitive and the latest version of the data (i.e. the one read after acquiring a lock).

Therefore, for each relaxed read operation, the application specifies (as a parameter of the *rlxread* primitive) an upper bound on the difference between the latest version and the one returned by the *rlxread* primitive call.

3.2 Controlling data freshness

Specifying the difference between the latest version and the one returned by the *rlxread* primitive is not a trivial problem. The hierarchical aspect of the data consistency protocol does not provide the ability to retrieve the latest version in one step. For a given data, different LDGs may store different versions indeed. The LDG that owns the lock associated to a given data hosts the lastest version of this data while the over ones may host an older one (as LDGs do not necessarily propagate every data update to the GDG). Furthermore, even client nodes attached to a same LDG may host different versions of a given data according to the last time they access this data: the data stored by a client node is only updated when it accesses the data (using the consistency protocol primitives).

To express the difference between the latest version and the version returned by the *rlxread* primitive, we introduce two parameters that take into account the two layers of the hierarchical consistency protocol.

- The D parameter is a constant attached to each piece of data.
- The w parameter (also called *reading window*) is specified for each call to the *rlxread* primitive.

The D constant corresponds to the number of times a LDG can give the exclusive lock to attached client nodes without sending updates to the GDG. The D parameter is set when the data is allocated by the service. Setting D to a small value forces the LDG to spread updates frequently, offering the possibility to get fresher data from the other LDGs. However, this solution adds an overhead due to GDG updates (releasing the lock, sending update messages, etc.). On the other hand, using a greater value let the writers performing writes within the LDG, without wasting time in frequent GDG updates. The counterpart is that the data versions returned by the relaxed read in other LDGs may be a bit older. For instance, if $D = 0$ LDGs have to spread their modifications to the GDG after each release of the exclusive lock by a client. In this case, all LDGs have the same version of the data (the latest). The D parameter has been inspired by the hierarchical synchronization protocol described in [7].

The w parameter is the *reading window*. It is specified for each call of the *rlxread* primitive. It defines an upper bound on the distance between the latest version of the data and the version returned by the relaxed read. Therefore, w must be greater than or equal to D. Considering the smallest value for w (i.e. $w = D$) implies that the relaxed read returns the LDG's version. This solution offers fresher data but it also implies more network traffic when data updates occur frequently (and therefore less efficient relaxed reads). Relaxing the read (i.e. using a greater value for w), enhances the observer access speed by reducing the network traffic but the relaxed read primitive may return older versions of the data.

Note that distances D and w are positive or null and w must be greater than or equal to D. The difference $w - D$ indicates the upper bound between the version of the data stored on the client's LDG and the one returned by the relaxed read primitive on the client's node. For instance, if $D = 3$ then all the LDG can successively give the lock up to 3 times without updating the GDG. If $w = 4$ then the version of the data read by the client is either the LDG's version of the data or the previous version.

For a given data, if a client has V_C as data version and if V_{LDG} is the version stored on its LDG, the client can use its own version V_C as long as the following condition is satisfied (α):

$$V_c \geq V_{LDG} - (w - D)$$

This condition is checked by the LDG each time a client node performs a relaxed read.

Efficient visualization relies on the correct tuning of both D and w parameters. Therefore, a smart combination of D and w parameters has to be used depending on the type of application that is monitored and the visualization accuracy that is required.

3.3 Example

Figure 3 illustrates the roles played by w and D within the hierarchical architecture of the protocol. The d data is available in 3 different versions stored on client nodes or LDGs (V_a in one cluster and V_b and V_c in a second one). Several clients acquire the lock, write the data, release the lock and send updates to LDG A, increasing the V_a version (1). Every D_d lock releases within LDG A, data updates are sent to the other LDGs (i.e. to the GDG) (2). At the same time, in the second cluster, Client C performs relaxed reads, using a window w as a parameter of each access. Therefore, a relaxed read request is sent from Client C to LDG B. This request contains 2 object 1) the w parameter and 2) V_c: the version of the data owned by client C (3). Depending on the evaluation of the α condition, the LDG B sends back either its V_b version of the data or a message that allows the client to use its own version (4).

3.4 Discussion

The relaxed read proposes an *extension* of the consistency model. Entry consistency is still preserved and guarantees that clients read an up-to-date version of the data, provided they acquire the associated lock. Besides, the entry consistency model is extended by a new feature: some controls are now available when processing a read without acquiring the lock.

Note that setting $D = 0$ and $w = 0$ is not equivalent to the classic sequence of performing a read after getting a read-lock. First, during the relaxed read, the lock can be acquired by another client which can modify the data. This is not allowed in the original entry consistency model. Second, between the moment when the LDG sends the data to the client and the moment when the data is returned by the *rlxread* primitive, new versions can be produced (as the protocol allows writes to continue). Therefore, the user has to know that this approach does not offer *strict* guarantees on data freshness. Providing more guarantees would require that the LDG wait for a client acknowledgment before accepting new updates. Such an approach would however be less efficient. Furthermore, these guarantees are not necessarily needed for the problem of efficient visualization within code-coupling applications.

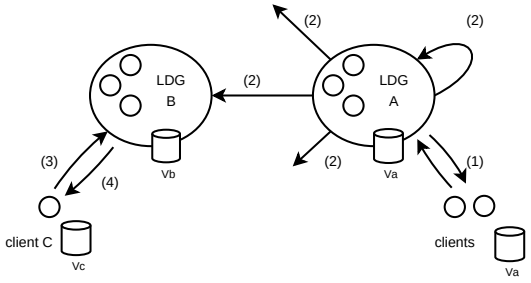


FIG. 3 – A relaxed read overview.

4 Preliminary evaluation

Besides JuxMem has been designed for grid, which may involve thousands of nodes, to illustrate the benefit of our consistency model extension, we run experiments with one single piece of replicated data (therefore using fewer nodes).

For all the experiments, we used the Grid'5000 platform [2], which gathers 9 clusters geographically distributed in several cities of France. These clusters are connected together through the Renater Education and Research National Network (1 Gb/s). For these preliminary experiments, we used from 9 to 25 nodes in 3 of these cities (Orsay, Rennes and Toulouse). In these 3 cities, nodes are connected through a Giga Ethernet network (1 Gb/s).

4.1 A visualization scenario

We consider a synthetic application running across 2 clusters (Rennes and Toulouse). As illustrated by Figure 4, Rennes's cluster contains processes performing writes on the shared piece of data, called *writers*. Processes performing reads called *readers* are located in Toulouse's cluster. A third cluster, Orsay's cluster, is used to visualize the application progress.

The experiments are configured as follow: the *writers* perform 50 writes each while the *readers* perform 50 reads each for this piece of data. The visualization process (the *observer* on Figure 4) performs 50 observations of the piece of data. Note that the data is replicated: there is one copy in each cluster. Each node hosting a copy is a LDG, the 3 LDG compose the GDG for this data.

The goal of these experiments is to evaluate the benefit of the consistency model extension upon the visualization process. Therefore, each test is performed twice depending on the visualization process: 1) using the *acquireRead* primitive (called *acquireRead-based*

visualization thereafter) , 2) using the *rlxread* primitive described in this paper. We also experiment relaxing the visualization modifying w and D parameters.

In order to evaluate the impact of the data size, 4 different sizes are experimented (1 KB, 512 KB, 1 MB and 10 MB).

Initially, we use a single *writer* and a single *reader*. Then, in order to vary the communication patterns the number of *writers* and *readers* is gradually increased up to 18 (performing $9 * 50$ writes and $9 * 50$ reads).

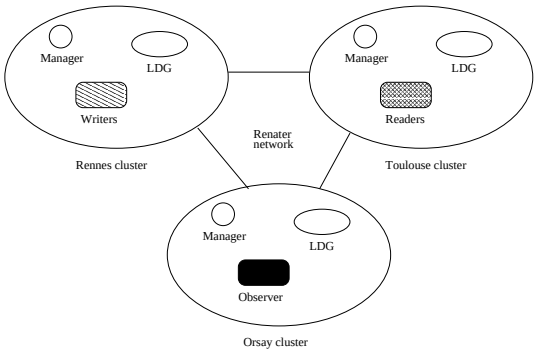


FIG. 4 – Experiments configuration

4.2 Results analysis

4.2.1 Benefi ts of the extension

the goal of these first set of experiments is to evaluate the benefit of the protocol extension even when parameters D and w are set to 0. As explain in section 3.4, this is not equivalent to read the data through the *acquireRead* primitive as no lock is acquired indeed.

Figure 5 shows the benefit for the visualization process. The improvement of approximately 80% is mainly explained by the fact that the visualization does not need to wait for a lock. The benefit is growing with the data size: larger the data is, longer the time to update the data and release the lock is. The benefit even reaches 94% for a 10MB piece of data (not displayed on the figure for readability reasons).

The visualization process is not the only one to take advantage of the *rlxread* primitive. The application itself shows a little improvement as it no longer has to wait for the visualization process to release its lock. Figures 6(a) and 6(b) respectively illustrate the benefit for the writer and the reader. However, the improvement is low because in the case of the *acquireRead-based* visualization, the read lock was already shared between the reader and the visualization process.

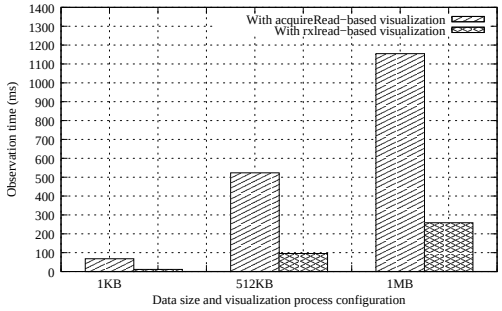


FIG. 5 – Observation improvement

Hence, the main benefit is for the visualization process, as shown by Figure 6(c) which summarizes the benefits for the reader, the writer and the visualization.

4.2.2 D and w influence

in order to evaluate the impact of the D and w parameters upon the visualization and the application, we run a second set of experiments, setting $D = 2$ and $w = 3$.

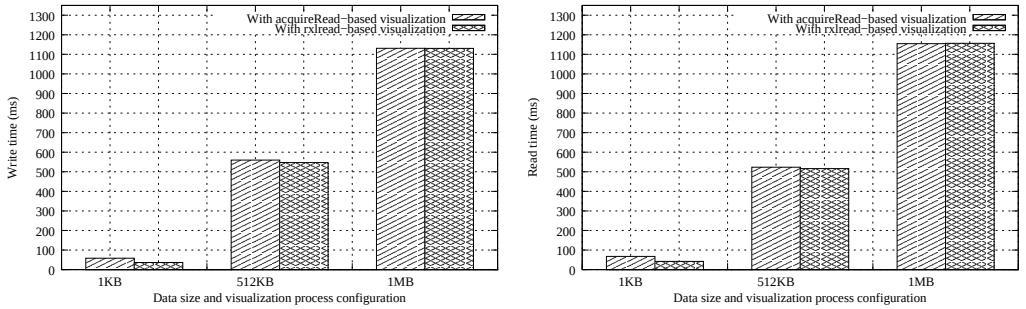
According to these values: 1) the LDG located in the Rennes' cluster propagates updates at least every 3 writes; 2) the LDG in Orsay's cluster sends back the data to the observer only if the difference between its version and the observer's one is more than 1 ($w - D$).

Figure 7 shows that relaxing constraint upon the data freshness results in an improvement for the visualization (33% for a data size of 1MB). Setting $w = 3$ reduces the probability that the observer needs to transfer the data. Therefore the improvement increases with the data size. On the over hand, the data returned by the *rlxread* primitive is a little bit less up-to-date.

The impact on the application is really low (almost null), as shown by figures 8(a) and 8(b).

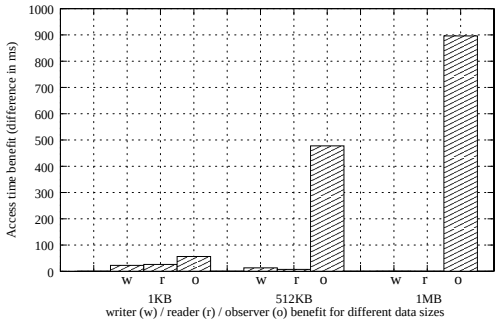
4.2.3 Varying communication patterns

Finally the number of writers in Rennes' cluster and the number of readers in Toulouse's cluster is increased in order to study what happens when stressing the protocol. Each test is run with both the *acquireRead-based* visualization (using the *acquireRead* primitive) and with the *rlxread-based* visualization. The size of the data is 1 KB. The results presented in Figure 9(a) show that the latency of the *rlxread* primitive is constant (and lower than the one of the *acquireRead-based* visualization) whatever the number of writers/readers is. The



(a)

(b)



(c)

FIG. 6 – Benefits of the extension. (a) Writer improvement. (b) Reader improvement. (c) Overall benefit.

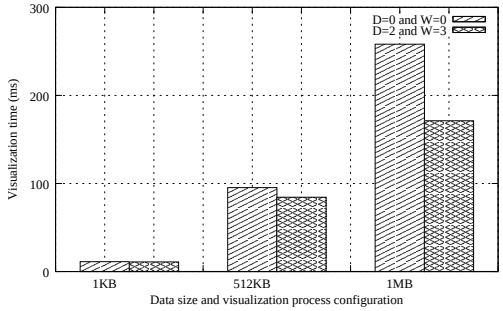
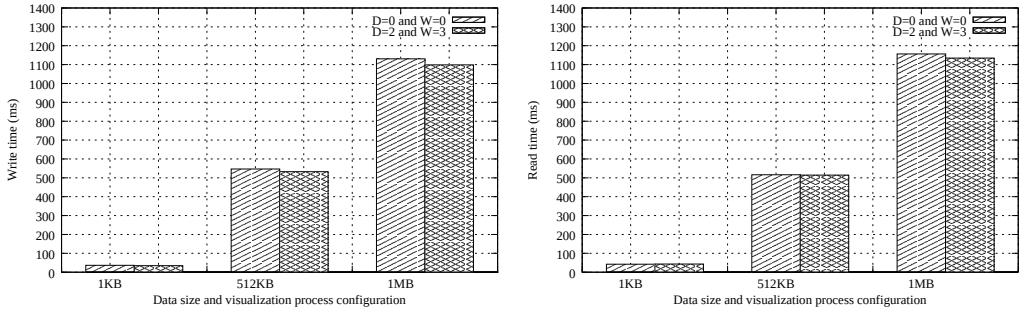


FIG. 7 – Observation improvement (D=2 W=3)



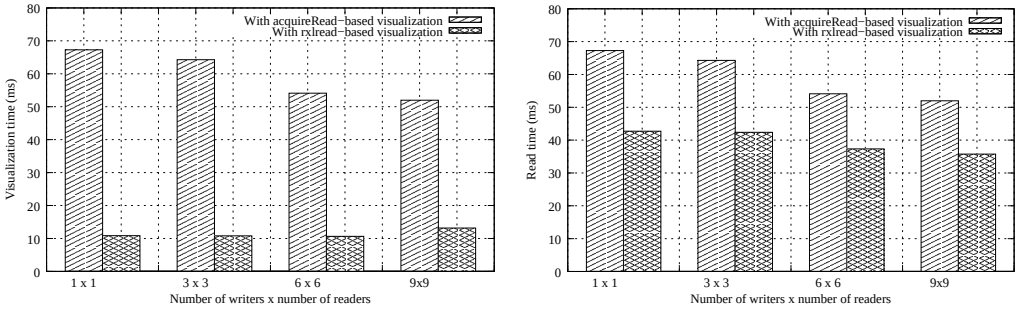
(a) (b)

FIG. 8 – D and w influence. (a) Writer improvement (D=2 W=3). (b) Reader improvement (D=2 W=3).

rlxread primitive induces communications between the visualization process and its LDG only.

The latency of the *acquireRead-based* visualization decreases while the number of readers increases: a high number of readers increases the probability that a read lock as already been given in the system. In this case, there is no need to wait for a release, the read lock can be shared by the numerous readers, providing a lower read latency.

However, as the number of writers and readers increases, the average write time grows. As the write lock is exclusive, the probability to wait for a release increases with the number of processes accessing the data with a lock (i.e. except the ones using the *rlxread* primitive). However, Figures 9(b) and 10 show that using the *rlxread* primitive provides a great improvement even with numerous writers and readers.



(a) (b)

FIG. 9 – Varying communication patterns. (a) Observation improvement. (b) Reader improvement.

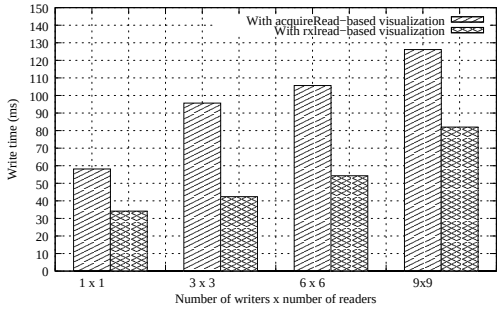


FIG. 10 – Writer improvement

As for the *acquireRead-based* visualization, the latency of the read operation decreases while the number of readers increases. There again the improvement offered by the *rlxread* primitive is significant.

5 Conclusion

Visualizing code coupling applications while they are running is a nice feature that may help to tune the application dynamically, to get preliminary results, to perform demos, etc. In this paper, we presented an extension of the entry consistency model that enables efficient relaxed reads concurrently to the application reads and writes. This provides the ability to perform an efficient, and still rather accurate visualization.

Preliminary results, obtained on the Grid'5000 testbed, show that using the new operation (*rlxread*) is a lot more efficient and slightly less intrusive than using *acquireRead* operation provided by the entry consistency model. The data version returned by the *rlxread* operation is not necessarily the most recent, however its "freshness" can be controlled and should be sufficient for visualization purposes.

We plan to further develop the extension proposed in this paper. The w parameter may be only a hint (e.g. *not accurate*, *accurate* or *very accurate*) according to the needs of the visualization process. JuxMem may then automatically decide what exactly the w parameter should be (which expresses the "freshness degree"), by taking into account parameters like the network load or the data update rate.

Références

[1] The JXTA (juxtapose) project. `http://www.jxta.org`.

[2] Projet Grid'5000. `http://www.grid5000.org`.

[3] Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Ian Foster, Carl Kesselman, Sam Meder, Veronika Nefedova, Darcy Quesnel, and Steven Tuecke. Data management and transfer in high-performance computational grid environments. *Parallel Comput.*, 28(5):749–771, 2002.

[4] Cristiana Amza, Alan L. Cox, Sandhya Dwarkadas, Pete Keleher, Honghui Lu, Ramakrishnan Rajamony, Weimin Yu, and Willy Zwaenepoel. TreadMarks: Shared memory computing on networks of workstations. *IEEE Computer*, 29(2):18–28, February 1996.

[5] Gabriel Antoniu, Luc Bougé, and Mathieu Jan. JuxMem: An adaptive supportive platform for data sharing on the grid. *Scalable Computing: Practice and Experience*, 6(3):45–55, November 2005. Extended version to appear in Kluwer Journal of Supercomputing.

[6] Gabriel Antoniu, Luc Bougé, and Mathieu Jan. JuxMem: Weaving together the P2P and DSM paradigms to enable a Grid Data-sharing Service. *Kluwer Journal of Supercomputing*, 2005. To appear. Preliminary electronic version available as INRIA Research Report RR-5082.

[7] Gabriel Antoniu, Luc Bougé, and Sébastien Lacour. Making a dsm consistency protocol hierarchy-aware: an effi cient synchronization scheme. In *Proc. Workshop on Distributed Shared Memory on Clusters (DSM 2003)*, pages 516–523, Tokyo, May 2003. Held in conjunction with the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID 2003), IEEE TFCC.

[8] Gabriel Antoniu, Jean-François Deverge, and Sébastien Monnet. How to bring together fault tolerance and data consistency to enable grid data sharing. *Concurrency and Computation: Practice and Experience*, (17), September 2006. To appear.

[9] Alessandro Bassi, Micah Beck, Graham Fagg, Terry Moore, James S. Plank, Martin Swany, and Rich Wolski. The internet backplane protocol: A study in resource sharing. In *CCGRID*

'02: Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, page 194, Washington, DC, USA, 2002. IEEE Computer Society.

[10] Brian N. Bershad, Mattew J. Zekauskas, and Wayne A. Sawdon. The Midway distributed shared memory system. In *Proceedings of the 38th IEEE International Computer Conference (COMPCON Spring '93)*, pages 528–537, Los Alamitos, CA, February 1993.

[11] John B. Carter, John K. Bennett, and Willy Zwaenepoel. Implementation and performance of Munin. In *13th ACM Symposium on Operating Systems Principles (SOSP)*, pages 152–164, Pacifi c Grove, CA, October 1991.

[12] Ian Foster, Carl Kesselman, and Steven Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *Supercomputer Applications*, 15(3):200–222, March 2001.

[13] Kai Li and Paul Hudak. Memory coherence in shared virtual memory systems. *ACM Transactions on Computer Systems*, 7(4):321–359, November 1989.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifi que,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399